Abstract

Machine learning (ML) is an increasingly popular approach/technique for analysing "Big Data" and predicting risk behaviours and psychological problems. However, few published critiques of ML as an approach currently exist. We discuss some fundamental cautions and concerns with ML that are relevant when attempting to predict all clinical and forensic risk behaviours (risk to self, risk to others, risk from others) and mental health problems. We hope to provoke a healthy scientific debate to ensure that ML's potential is realized and to highlight issues and directions for future risk prediction, assessment, management, and prevention research. ML, by definition, does not require the model to be specified by the researcher. This is both its key strength and its key weakness. We argue that it is critical that the ML algorithm (the model/s) and the results are both presented and that ML needs to be become machine-assisted learning like other statistical techniques; otherwise we run the risk of becoming slaves to our machines. Emerging evidence potentially challenges the superiority of ML over other approaches and we argue that ML's complexity significantly limits its clinical utility. Based on the available evidence, we believe that researchers and clinicians should emphasize identifying, understanding, and explaining (formulating) individual clinical needs and risks and providing individualized management and treatment plans, rather than trying to predict, or putting too much trust in predictions that will inevitably be wrong some of the time (and we do not know when).

Keywords: risk, prediction, suicide, machine learning, self-injury

Public health significance statement

Machine learning is a statistical approach/technique that is increasingly being used in an attempt to improve the accuracy with which risky behaviours and mental health problems are predicted. This study discusses some key considerations for using machine learning and making it even more useful.

Cautions, concerns, and future directions for using machine learning in relation to mental health problems and clinical and forensic risks: A brief comment on "Model complexity improves the prediction of nonsuicidal self-injury" (Fox et el., 2019)

We read the machine learning (ML) study by Fox et al. (2019) with great interest. It has many strengths. A number of potentially important and interesting predictors were measured in a fairly large sample of high-risk individuals four times over one month, with high retention rates. Measurement over four time points may have increased power to detect effects in these traditionally low base rate behaviours (it would in a regression model). The development of an evidence-base regarding the short-term prediction of nonsuicidal self-injury (NSSI), attempting suicide, and other types of risk is critical for clinical practice because clinicians are tasked with predicting current and imminent risk and clinical need and forecasting whether these things will change in the coming hours, days, and weeks.

ML is a popular approach/technique for analysing "Big Data" (Jordan & Mitchell, 2015) and increasingly being touted (e.g., Franklin et al., 2017) as one of the key solutions to improving the prediction of mental health problems and risks to self (NSSI, suicide attempts). We welcome innovation and efforts to model the complexity of human cognition and behaviour in order to improve understanding, prediction, assessment, management, and prevention. However, we are aware of few published critiques of ML as applied to the prediction of risk behaviours or mental health problems. We did find an article cautioning that "when one repeatedly searches a large database with powerful algorithms, it is all too easy to "find" a phenomenon or pattern that looks impressive, even when there is nothing to discover" in an article in the journal *Data Mining and Knowledge Discovery* (Salzberg , 1999, p. 1). No

methodology or statistical technique is perfect; a healthy scepticism is the best way to ensure that new methods are used thoughtfully, to realize their potential, and to avoid making conclusions that go beyond the data.

ML is a promising approach. It combines numerous variables to make very complex representations of data to improve prediction. This is potentially valuable because the available evidence indicates that clinical and forensic risks are extremely difficult to predict with sensitivity and specificity by clinicians or researchers (e.g., Fox et al., 2015; Franklin et al., 2017; Hanson & Morton-Bourgon, 2009; Messing & Thaller, 2013); probably because risk thoughts and behaviour are complex, multiply determined, relatively acute, and, most importantly, have a low base rate. We heartily welcome sophisticated efforts to predict and/or explain when and why clinical risks become apparent or change (Siddaway, Wood, O'Carroll & O'Connor, 2019a). However, the study by Fox et al (2019) and ML as an approach have some important limitations that need to be borne in mind as this research base evolves. This Brief Comment highlights a number of fundamental cautions and concerns with ML that are relevant when attempting to predict any clinical and forensic risk or mental health problem.

Limited generalisability. Although Fox et al (2019) report impressive predictive accuracy, their results are possibly not generalisable and may not be replicable as they were obtained using a very high-risk sample under very specific conditions (for a discussion of determinants of replication, see Siddaway, Wood, & Hedges, 2019b). Participants took part in an online survey, were each paid up to $70 for participating, and retention rates were unusually high. 88% of participants had a history of NSSI, 62% had a history of attempting suicide, and rates of NSSI were 20%, 35% and 41% at the three follow up time points. Predicting under these circumstances

(even in a crisis service) is very different to predicting in everyday clinical practice and at a population level.

It is noteworthy that few to no published studies have tested the replicability of ML algorithms on independent datasets. Indeed, in a recent paper outlining future directions for the suicide attempt literature and for research on psychological problems in general, it has previously been argued that ML models struggle to replicate in new samples (Franklin, 2019). Replication is always important (Siddaway et al., 2019b); however, given the concerns outlined in this Brief Comment, ML is an area where replication seems fundamental.

Methodological and reporting observations. The results presented by Fox et al. (2019) appear to validate the predictive validity of ML. ML substantially improved prediction beyond univariate and multivariate regressions. However, the authors did not, for example, test whether ML outperforms a latent variable multivariate regression model (which removes measurement error) or a multivariate regression model that included interactions and polynomials. More importantly, the clinical and research implications of the Fox et al. (2019) study (and other ML studies) are limited because the ML model/s that were used to obtain the results are not presented. It would be helpful to present the ML algorithm in detail, preferably in a way that is interpretable to clinicians. There are a number of considerations for studies of this kind. Were linear, nonlinear, and/or interaction effects modelled? How and for which individuals, using which variables? Were different models specified for different subpopulations? What were those models and what were the characteristics of the subpopulations?

For results to be useful, clinicians need to understand how the science informs their practice. They mainly require if-then information. For example, if I incorporate

X, Y and Z variables into my clinical assessment, then I will be more likely to be able to accurately and reliably predict and explain my patient's current and future risk to themselves. Or: For X subpopulation (e.g., males aged 20-30 with no history of NSSI), the risk of Y is Z. Transparency in ML model/s (i.e. *how* the model/s achieved the reported predictive accuracy) is necessary for establishing replication and would provide a "recipe" for which variables, in which combinations, for whom, when, provide accurate prediction.

Machine learning, by definition, does not require the model to be specified by the researcher. This is both its key strength and its key weakness. ML departs from the usual convention in psychological science that involves painstakingly specifying and describing statistical model/s, usually based on a theory, and presenting results in detail – practices that become increasingly important as the complexity of the model/s increases. Because of this long-established best practice, there is a vast methodological literature investigating how and why to specify particular statistical models in particular ways, given particular sets of circumstances and assumptions, to achieve particular goals. For ML as an approach to have sustained and robust value for researchers, clinicians, and policy makers, the ML algorithm (the model/s) *and* the results both need to be presented (cf. risk of bias for prediction model studies; Wolff et al., 2019). ML needs to become *machine-assisted* learning like other statistical techniques; otherwise we run the risk of becoming slaves to our machines.

Is the loss of interpretability worth it? Fox et al (2019) present impressive predictive accuracy – but at the cost of significantly increased and probably unusable complexity and compromised clinical utility. It is unclear whether ML's complexity is useful or necessary. In most psychological studies, main effects explain most of the variance and interaction effects do not substantially improve prediction, meaning that

complexity only slightly improves prediction, while interpretability is lost. It is difficult to imagine how very, very complex representations of data can be readily understandable to humans (e.g., enumerable slopes and intercepts) or have relevance to everyday clinical practice. Again, ML's attempt to model complexity is both its key strength and its key limitation.

This is of course the challenge and trade-off that researchers face when attempting to develop convincing and powerful theoretical models because as models become more complex in order to increase explanatory power, their clinical utility reduces (see Dalgleish, 2004). Simpler models are more accessible and have greater clinical utility but offer reduced explanatory power. How this trade-off is resolved depends on the aims and needs of the person using the theory (Dalgleish, 2004). Multiple regression and structural equation modelling, for example, are more interpretable methods than ML because explained variance, effect sizes, and predictive accuracy metrics can potentially be computed and reported.

Emerging evidence in the suicide research field, for example, brings into question the superiority of ML over other approaches. A recent review reported that positive predictive values (PPVs) are mostly "extremely low," leading the authors to conclude that ML currently offers limited clinical utility (Belsher et al., 2019). By contrast, Fox et al. (2019) were able to achieve strong predictive accuracy, which was probably attributable to the specific characteristics of their sample and research design. Their results may not replicate in an independent sample.

We are also aware of a forthcoming study that compared five different ML techniques to predict suicide ideation and attempts over one year using twenty different psychological constructs and found that unregulated multivariate logistic regression performed as well as ML (Van Mens et al., in press). Clearly, further

research is needed to clarify whether and under what conditions ML does or does not outperform other statistical approaches.

Limited clinical utility? It remains to be seen whether the potentially improved predictive accuracy that ML may offer actually improves clinical outcomes. ML could arguably usefully inform clinical practice even if the models are not described or comprehensible. Randomised controlled trials are required to test this question. We are aware of one fascinating effort in this regard (Jaroszewski, Morris & Nock, 2019); more are required, particularly regarding everyday clinical practice. Incorporating ML into everyday practice might make no difference or could even make things worse if it was somehow used as a substitute for a thorough, collaborative, theory and research-informed assessment and individualised risk formulation which describes, explains, and predicts risk and informs an individualised management plan.

Completely accurate and reliable prediction of mental health phenomena that have low base rates is probably impossible, so it is possible that ML is trying to achieve an impossible goal. A combination of relatively weak relationships between risk factors and risk behaviours such as NSSI (Fox et al., 2015) or suicide attempts (Franklin et al., 2017), combined with the low base rate of clinical and forensic risks, places a ceiling on PPVs and inevitably leads to false positives and false negatives. Fox et al. (2019) present a PPV of 94% at three days. Even with this high-risk sample and extraordinary predictive accuracy, the ML model presented by Fox et al (2019) still missed people at each time point (46 people at three days, 55 people at 14 days, and 53 people at 28 days).

It is worth remembering that predictive performance metrics (e.g., positive/ negative predictive values, area under the receiver operating characteristic curve [AUC], sensitivity, specificity) are dependent on the population and methodological

robustness of the study; transferring results to different settings may therefore be challenging as predictive values are affected by prevalence (Quinlivan et al., 2016). Evaluating performance metrics also depends on how the tools will be used in clinical practice (Quinlivan et al., 2016). For example, AUC, a measure of global predictive accuracy, is useful for comparisons and meta-analyses but less appropriate for clinical practice due to the lack of information on sensitivity (the proportion of people who repeat a risk behaviour and who are identified by a scale as being "high risk") and specificity (the proportion of people who do not repeat a risk behaviour and who are identified by a scale as being "low risk"). Sensitivity and specificity provide information about the performance of a particular scale compared to a reference standard (outcome or gold standard) but not the actual probability of the event occurring in practice. PPVs report the probability that a person identified as "high risk" actually goes on to repeat a risk behaviour, which may be more useful for clinicians.

It goes without saying that there are serious, potentially adverse consequences associated with over or under-estimating (predicting) clinical and forensic risks. "High"/"low" risk predictions in research and clinical practice are often wrong. For example, almost half of all the people who die by suicide come from the "low" risk strata (Large et al., 2017). The potential downside of risk stratification in clinical practice is that some people are incorrectly classified as being "high-risk" (false positives) and directed to unnecessary treatment whilst many others are classified as "low-risk" (false negatives) and denied much needed treatment (Large et al., 2017).

Conclusion. Researchers and clinicians have long wrestled with the daunting challenge of accurately and reliably predicting and managing different clinical and forensic risks and a very broad range of risk assessment tools, methodologies, and

statistical techniques have been developed to aid this endeavour. ML is an increasingly popular approach/technique that seems to have great potential. However, it is not a panacea – nothing is. Like all statistical techniques, it should be viewed as one potential tool that may be thoughtfully drawn upon and implemented to answer particular research questions and perhaps aid clinical practice. The parameters of ML's effectiveness (how, when, and why it might aid clinical practice) are empirical questions.

It is noteworthy that different types of clinical and forensic risk (risk to self, risk to others, risk from others) continue to be studied in isolation, with largely separate literatures. Our view is that many causal mechanisms and useful clinical principles and techniques may be applicable to the assessment, formulation, and management of all types of risk and that is therefore much to be gained by integrating different risk literatures.

Based on the available evidence across different literatures, our view is that researchers and clinicians who work with all types of clinical and forensic risk should emphasise identifying, understanding, and explaining (formulating) individual clinical needs and risks and providing individualised management and treatment plans, rather than putting too much trust in predictions that will inevitably be wrong some of the time (and we do not know when). This is because different clinical and forensic risks are complex, multiply determined, relatively acute, and, most importantly, have a low base rate. These tasks will be best achieved through a considered and individualised combination of (1) psychological theory, (2) empirical evidence regarding static and dynamic causal risk factors (perhaps the output of ML research), and (3) experienced professional judgment and decision-making (opinion). A combination of (2) and (3) are often referred to as 'structured professional judgment' or 'structured clinical

judgement' in the forensic literature. Risk factors need to be combined for each individual person through formulation and clinical judgement, rather than in a predetermined or simplistic fashion such as creating a total score on a scale. This approach has obvious, close parallels with the individualised assessment, formulation, and interventions that are advocated in psychological therapy.

References

Belsher, B. E., Smolenski, D. J., Pruitt, L. D., Bush, N. E., Beech, E. H., Workman, D. E., Morgan, R. L., Evatt, D. P., Tucker, J., & Skopp, N. A. (2019). Prediction models for suicide attempts and deaths: A systematic review and simulation. *JAMA Psychiatry, 76,* 642-651.

Dalgleish, T. (2004). Cognitive theories of posttraumatic stress disorder: The evolution of multi-representational theorizing. *Psychological Bulletin, 130,* 228-260.

Fox, K. R., Franklin, J. C., Ribeiro, J. D., Kleiman, E. M., Bentley, K. H., & Nock, M. K. (2015). Meta-analysis of risk factors for nonsuicidal self-injury. *Clinical Psychology Review*, 42: 156-67.

Fox, K.R., Huang, X., Linthicum, K.P., Wang, S.B., Franklin, J.C., & Ribeiro, J.D. (2019). Model complexity improves the prediction of nonsuicidal self-injury. *Journal of Consulting and Clinical Psychology, 87,* 684-692.

Franklin, J.C. (2019). Psychological primitives can make sense of biopsychosocial factor complexity in psychopathology. *BMC Medicine*, 17, 187.

Franklin, J.C., Ribeiro, J.D., Fox, K.R., Bentley, K.H., Kleiman, E.M., Huang, X., Musacchio, K.M., Jaroszewski, A.C., Chang, B.P., & Nock, M.K. (2017). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin, 143,* 187-232.

Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment, 21,* 1-21.

Jaroszewski, A.C., Morris, R., & Nock, M.K. (2019). Randomized controlled trial of an online machine learning-driven risk assessment and intervention platform

for increasing the use of crisis services. *Journal of Consulting and Clinical Psychology, 87*, 370-379.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science, 349*, 255-260.

Large, M. M., Ryan, C. J., Carter, C., & Kapur, N. (2017). Can we usefully stratify patients according to suicide risk? *British Medical Journal, 359*, j4627.

Messing, J. T., & Thaller, J. (2013). The average predictive validity of intimate partner violence risk assessment instruments. *Journal of Interpersonal Violence, 28*, 1537–1558.

Quinlivan, L., Cooper, J., Davies, L., Hawton, K., Gunnell, D., & Kapur, N. (2016). Which are the most useful scales for predicting repeat self-harm? A systematic review evaluating risk scales using measures of diagnostic accuracy. *BMJ Open, 6*; e009297. doi:10.1136/bmjopen-2015-009297.

Quinlivan, L., Cooper, J., Meehan, D., Longson, D., Potokar, J., Hulme, T., Marsden, J., Brand, F., Lange, K., Riseborough, E., Page, L., Metcalfe, C., Davies, L., O'Connor, R., Hawton, K., Gunnell, D., & Kapur, N.. (2017). Predictive accuracy of risk scales following self-harm: Multicentre, prospective cohort study. *The British Journal of Psychiatry, 210*, 429–436.

Salzberg, S. L. (1999). On comparing classifiers: A critique of current research and methods. *Data Mining and Knowledge Discovery, 1*, 1-12.

Siddaway, A. P., Wood, A. M., & Hedges, L. V. (2019b). How to do a systematic review: A best practice guide to conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. *Annual Review of Psychology, 70*, 747-770.

Siddaway, A.P., Wood, A.E., O'Carroll, R.E., & O'Connor, R.C. (2019a). Characterizing Self-injurious Cognitions: Development and Validation of the Suicide Attempt Beliefs Scale (SABS) and the Nonsuicidal Self-injury Beliefs Scale (NSIBS). *Psychological Assessment, 31*, 592-608.

Van Mens, K., de Schepper, C. W. M., Wijnen, B., Koldijk, S., Schnack, H., de Looff, P., Lokkerbol, J., Wetherall, K., Cleare, S., O'Connor, R. C., & de Beurs, D. (2019, April 4). Predicting future suicidal behaviour with different machine learning techniques: A population-based longitudinal study. https://doi.org/10.17605/OSF.IO/35ATY.

Wolff, R.F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Annals of Internal Medicine, 170,* 51–58.